DOCUMENT RESUME

ED 036 190 24 EM 007 776

AUTHOR UNKS, NANCY J.; COX, RICHARD C.
TITLE A MODEL FOR THE EVALUATION OF A TESTING PROGRAM. WORKING PAPER NUMBER FOUR OF THE PROGRAM OF STUDIES IN EDUCATIONAL RESEARCH.
INSTITUTION RESEARCH FOR BETTER SCHOOLS, INC., PHILADELPHIA, PA.
SPONS AGENCY OFFICE OF EDUCATION (DHEW), WASHINGTON, D.C. BUREAU OF RESEARCH.
BUREAU NO BR-6-2867
PUB DATE 23 APR 68
CONTRACT OEC-1-7-062867-3053
NOTE 7P.; PAPER PRESENTED AT ANNUAL MEETING OF AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (CHICAGO, ILLINOIS, FEBRUARY 9, 1968)

EDRS PRICE EDRS PRICE MF-$0.25 HC-$0.45
DESCRIPTORS *EVALUATION METHODS, MEASUREMENT TECHNIQUES, *PROGRAM EVALUATION, TEST CONSTRUCTION, *TESTING PROGRAMS, TEST RELIABILITY, TEST VALIDITY
IDENTIFIERS INDIVIDUALLY PRESCRIBED INSTRUCTION, IPI

ABSTRACT

THE EVALUATION OF A TESTING PROGRAM IS NECESSARY BEFORE OR DURING A SOUND TOTAL PROJECT EVALUATION. IDEALLY, THE TESTING PROGRAM STUDY SHOULD BE CONCURRENT WITH, AND EQUAL IN MAGNITUDE TO, THE TOTAL PROJECT EVALUATION. STEP ONE IN AN EVALUATION IS TO DEFINE THE TESTING PROGRAM'S OBJECTIVES IN OPERATIONAL TERMS. STEP TWO IS A THOROUGH DESCRIPTION OF THE INNOVATION TO BE STUDIED. THEN THE EVALUATION PROGRAM SHOULD EXAMINE THE INSTRUMENTS USED TO CONDUCT TESTS, OBTAINING VALIDITY, RELIABILITY, AND ITEM ANALYSIS DATA FOR ALL SUCH INSTRUMENTS. A SUMMARY IS THEN MADE WHICH INTERPRETS THE INFORMATION ACCUMULATED IN THE FIRST THREE PHASES. A DIAGRAM OF SUCH AN EVALUATION PROCEDURE IS PRESENTED. (JY)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED036190

Working Paper No. 4

of the Program of Studies in Educational Research

April 23, 1968

A MODEL FOR THE EVALUATION OF A TESTING PROGRAM

Nancy J. Unks and Richard C. Cox

EM004 716

A Model for the Evaluation of a Testing Program[1]

Nancy J. Unks and Richard C. Cox

University of Pittsburgh

Well planned education and curriculum innovations include comprehensive evaluation activities as an integral part of the project. Evaluation in such programs implies measurement of outcomes on a number of dependent variables; this measurement implies testing in one form or another. Typically the most relevant and meaningful measurement devices that can be used for this evaluation are tests designed specifically for the curriculum innovation being studied. In a large scale program, the rather formidable task of producing these tests may be accomplished by a staff of test construction specialists--a sub-group within the larger project. The testing program designed by this test construction group provides for the assessment of pupil performance within the educational innovation. Since pupil performance is usually a major criterion for evaluating the entire project, an accurate assessment of that performance is essential to the project evaluation. In other words, the evaluation of an entire innovation is often dependent upon measurements made by a testing sub-program. If these measurements are not meaningful or reliable, then the evaluation may be subject to question. At this point a rationale for evaluating the testing sub-program becomes apparent. It is important for the researcher who is interested in the worth of the total project to first know the worth of his instruments. The evaluation of the testing program is a necessary pre- or co-requisite for a sound total project evaluation.

[1]Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, February 9, 1968.

Procedures for studying the testing program may include many of the principles and practices of evaluating educational programs in general. Ideally, the testing program study should be concurrent with and equal in magnitude to the total project evaluation. Figure 1 outlines a model or generalized plan for accomplishing the evaluation of a testing program. (This model has been adapted from a general evaluation model proposed by C. M. Lindvall at the University of Pittsburgh.) The line of small boxes at the top of the page represents a total project evaluation. The testing program study that is the topic of this paper parallels the total project study. The large boxes show the four major components or phases of the testing program evaluation. The arrows represent connecting links between phases: they represent questions the evaluator asks about the information he collects within the four components. Some of these questions are elaborated below the model. They are arbitrary, and they constitute the most subjective aspects of the evaluation, but the evaluator must attempt to answer them impartially and support his answers with objective evidence. The various procedures in the model and some details will now be elaborated.

The first phase on this procedural model is to define the testing program objectives. These objectives must be expressed in quite unambiguous, operational terms so that their achievement can be assessed. They must be consistent with the total project goals, and, in addition, they should elaborate the unique functions of the testing program and define its role in the project. It cannot be emphasized enough that the objectives must be operational. For example, it is not sufficient to say that an objective of the testing program is to write "good" achievement tests for the project. The type and content of these achievement tests must be specified, and criteria for what will be called a "good" test must be defined in terms

of validity, reliability, and item characteristics.

Essential to the evaluation of any program is a thorough description of the innovation to be studied. This is the second component of the model for the evaluation of a testing program. The evaluator should carefully study, observe, and define the written plan for, and the actual operation of the testing program. He should describe in detail its personnel, its facilities, and the instruments and measurements it produces, taking into account the relationship of the actual operation to the stated objectives.

The third component of the model is what might be considered the heart of the evaluation--the actual assessment of the testing program's outcomes. It has already been suggested that measurement of the achievement of the testing program objectives depends upon how the objectives themselves are stated. At this point a discussion of an existing testing program and the assessment of two of its objectives will help clarify this third component.

One of the projects of the University of Pittsburgh's Learning Research and Development Center is Individually Prescribed Instruction (IPI). The IPI system includes a testing sub-program which provides the diagnostic instruments necessary for measuring pupil achievement in reference to the IPI curriculum. In others words, it produces achievement tests which assess a pupil's mastery of specific skills in the curriculum. The first operational objective of the IPI testing program is to provide achievement tests which are specifically content-referenced to the behavioral objectives of the IPI curricula. To assess this goal, a check can be made as to whether such tests exist for each curriculum objective.

Another operational objective of the testing program is to place pupils in proper work levels to begin study in IPI at the beginning of each school year or when they first enter an IPI school. Assessing the achievement of this objective is a little more complicated. Not only is it necessary to check whether placement tests exist for all levels and units of work, but also it is essential to find out whether the tests place pupils in the proper work levels. An estimate of placement test validity can help give an idea of how accurately it assesses pupil achievement. A concurrent validity can be obtained by administering, to a selected sample of students, both the placement test and the set of pretests covering the same units of work. Then, results of the placement tests can be compared with those of the pretests which supposedly measure the same skills in greater detail. Another way to find out whether pupils have been properly placed is to examine their work patterns during the first two months of school and identify cases of misplacement. If a pupil seems to have unusual difficulty with the work, or if he goes through it with extreme ease, it may be an indication that he has been misplaced.

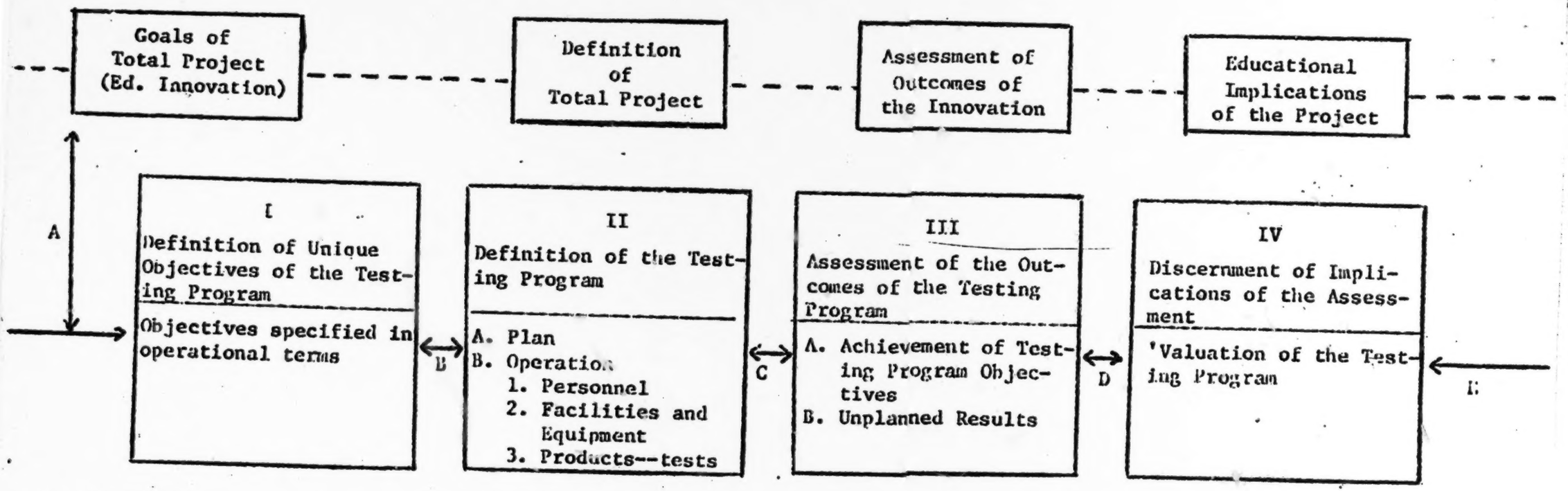
In general, a major portion of the assessment of any testing program consists of evaluating the instruments it produces. This means obtaining validity, reliability, and item analysis data for all such instruments and comparing this information with the standards established in the testing program objectives. For example, placement tests would be designed as general tests covering many skills and should, therefore, have low internal consistency reliabilities and low inter-item correlations. Tests of single skills, on the other hand, should be quite homogeneous and should have high internal consistency and high inter-item correlations.

If the testing program is large and produces many instruments, it is almost a necessity to use computer facilities to collect and organize the data as well as to provide statistical analyses. If such facilities are available and are being used in the total project evaluation, they should certainly be utilized in the testing program assessment; to use the computer for statistical analyses of test results without detailed analyses of test characteristics is rather meaningless.

Along with the objective assessment of the tests, the evaluator should also be concerned with the more subjective observation, description, and evaluation of unexpected or unplanned outcomes of the testing program. For example, he must be alert to notice the effect of delays in getting needed materials, changes in the goals of the total project, changes in the curriculum, or lack of communication between members of the testing staff and the rest of the project. All such observations should be recorded regularly and explicitly.

The forth phase of the model is the one in which the evaluator summarizes and interprets the information he has accumulated in the first three phases. He makes his interpretations in light of the objectives of the testing program and the goals of the total project. (Notice that the diagram of the model can be made into a cylinder so that arrow E becomes two directional, connecting the interpretation phase, IV, with phase I, the objectives of the testing program. Arrow A, then, indirectly relates the implications of the assessment to the goals of the total project.) In compiling the results, the evaluator attempts to establish the worth of the testing program--to place a 'valuation on it.' In conclusion, the rationale for evaluating the testing program can be expressed in one question--Do the measurements made by this testing program provide a sound basis for a total project evaluation? The answer to this question epitomizes the entire evaluation study.

Figure 1



1. Are testing program objectives relevant to goals of the total project?
2. Do the plan and operation of the testing program follow logically from its objectives?
3. Does operation suggest that changes are needed in the objectives to make them more specific and more readily implementable in practice?
4. Does the assessment suggest that changes in the plan and operation are needed in order to make the testing program function more efficiently?
5. Is the assessment comprehensive, covering all aspects of the testing program as defined in observation of its operation?
6. What does the assessment imply regarding the functions of the testing program and total project and the relationship of the testing program to more ultimate educational goals?
7. Can the implications drawn be justified logically by the empirical evidence?
8. Do outcomes show that the testing program objectives have been accomplished?
9. Do outcomes suggest needed changes in testing program objectives?
10. Does the testing program contribute to the achievement of total project goals?
11. Are the evaluation of the testing program and the results of that evaluation concomitant and consistent with the total project evaluation?
12. Do measurements made by this testing program provide a sound basis for a total project evaluation?